IRASS Journal of Multidisciplinary Studies
https://irasspublisher.com/journal-details/IRASSJMS
ISSN *(Online) 3049-0073*

The Product Manager's Role in AI Security: Preventing Data Leaks and Model Manipulation in Consumer Applications

Obianuju Gift Nwashili[1*], Kehinde Daniel Abiodun[2], Olamide Amosu[3], Sonia Oghoghorie[4]

[*1-2-3-4] Independent Researcher

Corresponding Author: Obianuju Gift Nwashili

Independent Researcher

Article History

Received: 16 / 10 / 2025

Accepted: 30 / 11 / 2025

Published: 10 / 12 /2025

Abstract: With the rapid adoption of artificial intelligence (AI) in consumer products, Product Managers (PMs) face an unprecedented responsibility: AI security. This article explores the critical role of PMs in identifying and mitigating two primary risks in AI systems: data leaks (such as potential exposure of sensitive training data through crafted prompts) and model manipulation (such as adversarial attacks that cause unintended system behaviors). We present a pragmatic, PM-centric framework for managing AI security risk that can be woven into existing product development workflows. First, PMs should facilitate threat modeling as part of the discovery process to identify potential misuse cases and inform the risk management strategy. Second, PMs can define security-oriented user stories and architectural guardrails during the design phase. Third, PMs should coordinate with security teams to perform red-teaming exercises before launch. Continuous prevention requires PMs to establish data governance as a top priority and promote consistent robustness testing practices. Success in this endeavor requires the PM to be the connective hub in the organization—translating technical risk to business risk and collaborating closely with cross-functional teams including Security, Legal, and Engineering to implement an effective security strategy. By building these elements into the fabric of how they work, PMs can position themselves as the first line of defense in upholding user trust and product integrity.

Keywords: *AI Security, Product Management, Data Leaks, Model Manipulation, Threat Modeling, Adversarial Attacks, Consumer Applications*.

How to Cite in APA format: Nwashili, O. G., Abiodun, K. D., Amosu, O. & Oghoghorie, O. (2025). The Product Manager's Role in AI Security: Preventing Data Leaks and Model Manipulation in Consumer Applications. *IRASS Journal of Multidisciplinary Studies, 2(12), 30-35.*

Introduction

The Security Challenge and Opportunity for the AI Product Manager

The rise of AI-enabled consumer products has redefined traditional cybersecurity models, creating a new and complex threat landscape (Isaac & Reno, 2023). Traditional security mechanisms are inadequate in the face of threats targeting data integrity, model robustness, and the potential for misuse or malicious manipulation (Tallam, 2025). AI introduces unique challenges, including data poisoning, model inversion, adversarial attacks, and more (Roshanaei et al., 2024; Tallam, 2025). These threats can compromise system functionality, expose sensitive information, or cause unintended and potentially harmful behaviors (Roshanaei et al., 2024; Tallam, 2025).

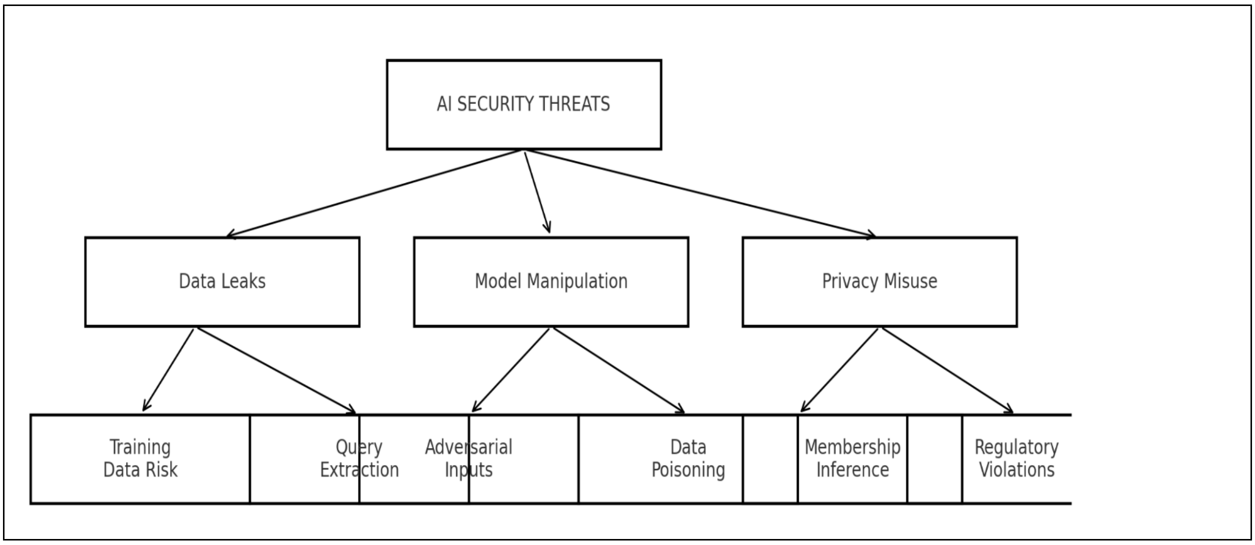

Figure 1: *Categorization of primary AI threat vectors that Product Managers must prioritize, showing how data leaks, model manipulation, and privacy misuse create distinct yet interconnected security risks in consumer AI products.*

The response to these vulnerabilities must be a shift in mindset towards a holistic and proactive security-first approach throughout the AI development lifecycle (Rangaraju, 2023; Tallam, 2025). Instead of reacting to threats, AI systems must be designed to anticipate, detect, and respond to potential security incidents. This proactive approach is at the heart of a new paradigm known as “AI Guardianship,” which envisions using AI-driven technologies like machine learning and predictive analytics to continuously monitor and respond to evolving threats. It allows for systems to detect anomalies and adapt their defenses, promoting an environment of continuous learning and improvement. This approach is a kind of “Secure by Intelligence” paradigm shift, in which the organization shifts from reactive to proactive defensive stances (Rangaraju, 2023).

For the Product Manager, this shift is not just a technical problem to solve. Instead, this new approach to security is the very foundation for Product Manager leadership. A proactive, security-first mindset is the only reliable way to build the trust of others in the AI ecosystem, with a focus on security, safety, and transparency built into the very foundations of a new product, from day one (Sidhpurwala et al., 2025). This is the best way to create strong, resilient, and ultimately trustworthy systems, from which all subsequent safety, transparency, and accountability efforts are more likely to succeed (Tallam, 2025).

Building robust security into AI systems is crucial due to the widely distributed AI infrastructure in the cloud, at the edge, or in some hybrid combination of these (Tallam, 2025). The larger attack surface across these three environments dramatically increases the opportunities for bad actors to target systems. The complexity of AI models, especially the deep neural networks (DNN) that underpin many AI systems, and the sheer volume of data these systems ingest and process, also represent new and largely unexplored areas of potential vulnerabilities (Tallam, 2025). Therefore, they must be addressed upfront, with a focus on security by design, and the PM will need to be central to the technical conversation, translating these risks into product strategy.

The PM-Led Security Framework

Integrating Safety into the Product Lifecycle

PMs must ensure the development and adoption of a repeatable framework that incorporates and operationalizes the security design into the product lifecycle. Such a framework helps the PM bridge the gap between the business requirements, user needs, and technical security requirements.

Table 1: *PM Security Responsibilities Across Product Lifecycle*

Product Stage	**Core PM Actions**	**Security Focus**	**Cross-Functional Partners**
Discovery	Lead threat modeling	Identify misuse risks	Security, Engineering
Design	Write security-first user stories	Guardrails & data boundaries	Legal, Architecture
Pre-Launch	Plan red-teaming	Robustness testing	Security Ops
Post-Launch	Monitor security metrics	Ongoing anomaly defense	DevOps, Data Science

Discovery & Planning: Threat Modeling: Initiate and lead a security threat modeling exercise during the discovery phase with the engineering and security teams. Brainstorm all possible attack scenarios unique to the AI use case, such as prompt injection to extract data from the system, generating adversarial examples to manipulate the model, or tricking the model into making biased decisions. This exercise will help to capture and prioritize the security requirements and inform the architectural foundation of the product.

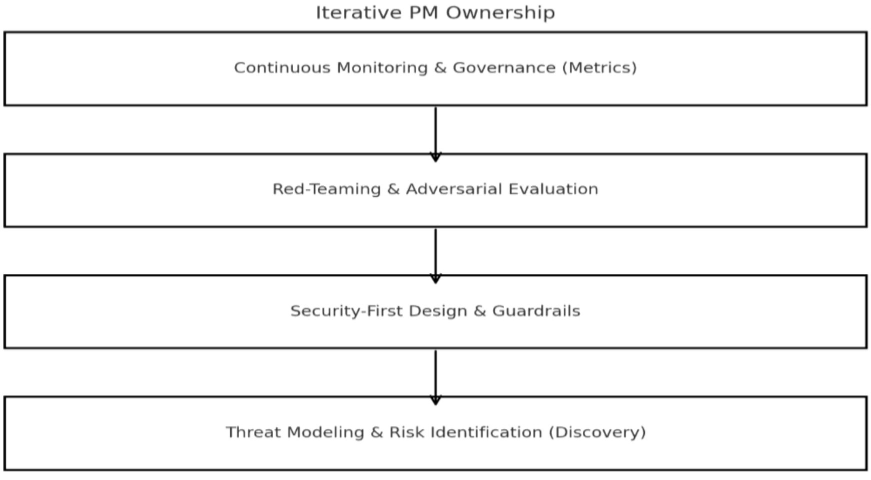
Iterative PM Ownership
Continuous Monitoring & Governance (Metrics)
Red-Teaming & Adversarial Evaluation
Security-First Design & Guardrails
Threat Modeling & Risk Identification (Discovery)

Figure 2: *PM-Led AI Security Framework*

Design & Specification: Creating "security-first" Stories and Acceptance Criteria: Leverage the identified threats to create "security-first" user stories with acceptance criteria to drive security-focused design and implementation. For example, a user story for ensuring security against prompt injection attacks could be: "As a system, I must sanitize and validate all user inputs to prevent prompt injection attacks that could lead to training data leakage." Additionally, architectural guardrails, such as mandatory input sanitization layers, output filters to prevent sensitive data leakage, and logging mechanisms for anomaly detection, should be established during this phase.

Pre-Launch: Red-Teaming & Security Metrics: In collaboration with dedicated security teams, conduct red-teaming exercises before launch. These are controlled adversarial simulations that attempt to exploit the system using real-world attack scenarios. Also, define AI-specific security metrics (e.g., mean time to detect an adversarial example, rate of blocked data extraction prompts) that will be monitored in production post-launch. This shifts security evaluation from a binary "pass/fail" to a measurable metric.

Mitigating Core AI Threats

Strategic Playbook for PMs

For the purpose of facilitating planning and resource allocation for mitigation action, Product Managers should narrow their efforts, influence, and ask-for-help scope and concentrate on the two key types of AI-specific attack vectors that are under active exploitation based on current threat intelligence (Roshanaei et al., 2024; Tallam, 2025).

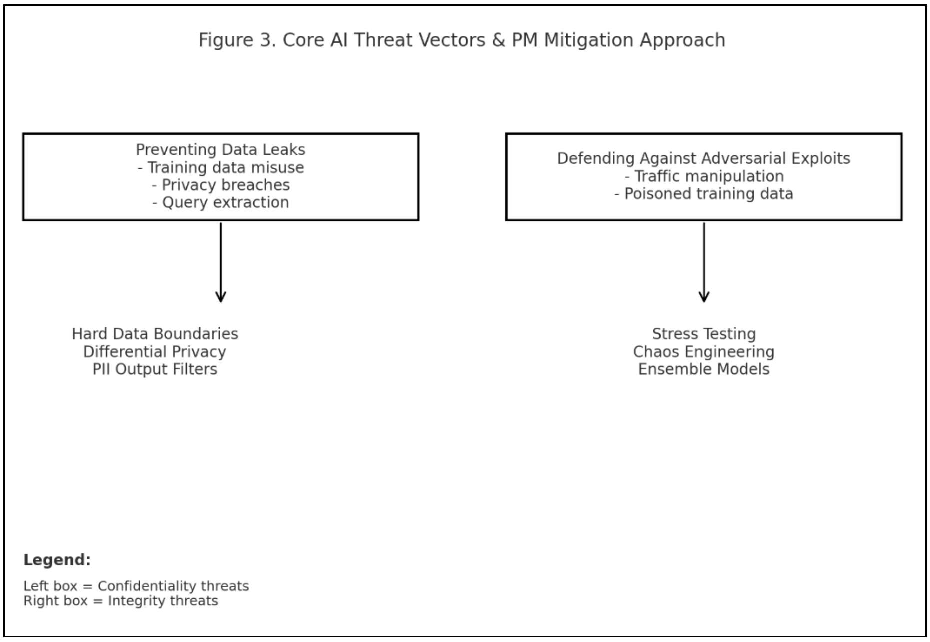

Figure 3: *Core AI Threat Vectors & PM Mitigation Approach*

Focus Area 1. Preventing Data Leaks

AI and machine learning based systems are data by design, with model training data and runtime request/response data being particularly likely sources of leakage risk. The attack is not as simple as direct database exfiltration (though this may still happen for weaker, early-stage applications in some areas) – it is usually a structured extraction process done by an attacker first reverse engineering the data sources by generating model queries based on a priori knowledge and then iteratively observing the model's outputs in order to infer and reconstruct actual training data records or membership (membership inference attacks) (Tallam, 2025). This could have serious PII breach, financial or health-related impacts for consumer-facing products that work with regulated or private data.

PM Playbook: For those two reasons, along with potentially long tail liability, the PM needs to ensure privacy by design and hard boundary enforcement are kept top of mind through all stages of product and engineering cycle, from the earliest scoping days on – that is, minimize training data risk as a baseline and have explicit definitions and hard limitations on the scope of user queries possible via the trained model's application UI.

Table 2: *Example Security-First User Stories*

User Story Type	Example	Acceptance Criteria
Input Validation	"As a system, I must sanitize all prompts."	All prompts filtered for unsafe patterns
PII Protection	"As a PM, I must prevent sensitive data output."	Automatic masking applied to names, IDs
Behavior Limits	"As a PM, I must block excessive query patterns."	Query length & cost thresholds enforced
Monitoring	"As a PM, I need anomaly alerts."	Alerts triggered for unusual traffic

* PM must be a strong advocate to place privacy controls as the primary design constraint on data decisions throughout the technical life cycle. This is achieved by PM first requiring engineers to prove that no existing or future PII or regulated data is

memorized or identifiable in the output in any capacity (or can be for even one user on the model) from technical controls that actively disassociate trained model outputs from the training data it was built on, for example by inserting data anonymization, synthetic data, or other forms of differential privacy into the training pipeline.

* Explicit Data Boundaries: The PM must also actively drive coordination with legal, compliance and engineering teams to map and understand precisely what data was used to train the model and enforce hard limits on operational side that limit user behavior, therefore disabling the attack vectors. This includes on the product side making specific API design decisions that dissuade or outright prevent model queries that could be interpreted as seeking to leverage data extraction (Tallam, 2025). Think hard caps on total query complexity or length, output filters for PII or other sensitive data, and logging of both suspiciously long or expensive query flows, as well as all queries over a defined high-volume threshold for potential manual review. These requirements must be described as hard non-negotiable product scope.

Focus Area 2: Defending Against Adversarial Exploits

The second focus area in the current threat landscape are model manipulation attacks. In this case, attack surfaces, vectors and motivations also differ from legacy systems. Adversarial attacks refer to carefully crafted data that is presented to an AI model in order to produce an unintended, biased or dangerous response (examples include often sub-visual perceptibility-misleading adversarial inputs (Roshanaei et al., 2024)). Data poisoning attacks usually happen at the training data stage and refer to injecting attack vectors into the training data itself with the aim of creating targeted vulnerabilities in the logic of trained model predictions (these can be also task-agnostic model evasion attacks where attackers simply hope to trigger widespread misclassification) (Roshanaei et al., 2024). These types of attacks can lead to direct product damage in functionality or performance terms, malicious disinformation spreading (or generating) and even system-wide process failures.

PM Playbook

Stress-testing AI Robustness Must be a Standard Product Feature: Ensure that adversarial attack surface and risk is identified and regular AI robustness testing is treated and scheduled in the dev-ops pipeline as an inherent part of the CI/CD (continuous integration/continuous deployment) process. This needs to include hard constraints, controlled as code and stored in pipeline-run processes, on when a model change/upgrade is blocked or must be rolled back based on accepted/unacceptable robustness results. This may also require reserving sprint time slots for “chaos engineering” activities where team generates a pool of adversarial and poisoning testing examples against current dev-stage model artifacts as an explicit block prior to further investment in sprint.

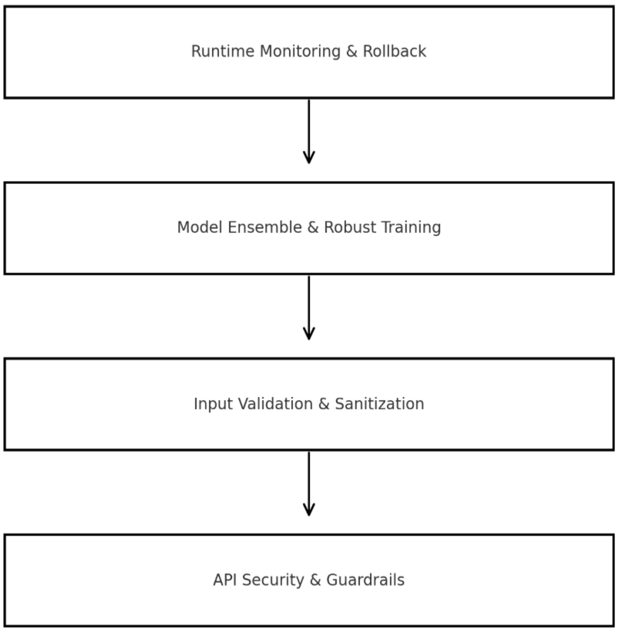
Runtime Monitoring & Rollback
Model Ensemble & Robust Training
Input Validation & Sanitization
API Security & Guardrails

Figure 4: *Layered Defense Architecture for ML Models*

Layered, Integrated Defense Architecture for ML Models: As with the data risk focus area, the PM should be an advocate to ensure the applied security architecture for the model is a best-in-class, layered one. This not only requires engineering process excellence in building layers (input validation/sanitization, robust API design) and will use but also specific PM push for the following hard, potentially architectural guardrails on technical side (see additional proposal in Appendix B that has links to open-source AI Guardianship examples for each point):

- Input Validation & Sanitization Layers: Must ensure and push for layers that can identify and stop known types of malicious input patterns before they can even hit core model layers for data poisoning types.
- Model Ensembles / Robust Training: Should support using ensembles of multiple models that are themselves either trained with one another in more robust/fault-tolerant way or at least can compare outputs for sanity in “model guardianship” concept.
- Automated Runtime Monitoring / Rollback Triggers: Must demand that a set of technical criteria be established for key metrics (e.g., aggregate certainty of predictions, total validation errors etc.) that cannot be breached and, in the event they are, can trigger both automated warnings and potential rollback process flows on an urgent, automatic basis.

Discussion

Balancing Innovation with Impermeable Security

The primary concern threaded throughout the article, which is the paradox of embedding stringent AI security in dynamic product cycles, is fundamentally a leadership and cultural challenge as much as a technical one. The need for the Product Manager to function as the fulcrum balancing these competing demands is both undeniable and a recipe for inherent conflict. The push to innovate and release features is contrasted starkly with the need to construct systems that can resist new, evolving attacks that threaten not just user safety but also erode user trust and the brand itself (Isaac & Reno, 2023).

Specifically, this case postulates several instructive points and implications:

Security as a Value Proposition, Not a Tax

As consumers become more aware and regulations more stringent, security is morphing into not just a baseline expectation but a key value driver in its own right. A “secure by design” product in an increasingly AI-infused marketplace is better positioned to build trust and therefore retain customers and stake market share (Rangaraju, 2023). The PM who can authentically and transparently communicate this aspect of the product—not just in features but in processes like transparency reports, clear data usage policies, and control mechanisms—will turn a necessity into a competitive edge.

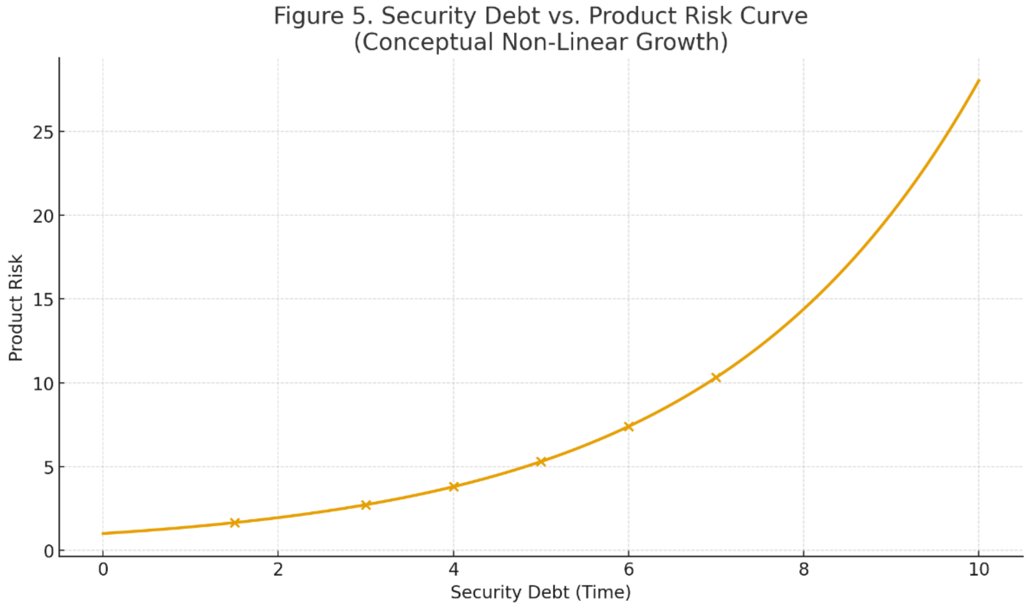

Figure 5: *Security Debt vs. Product Risk Curve*

Technical Debt Reimagined

For AI systems, technical debt acquires a new and more sinister form. An unaddressed security bug in a model, its data pipeline, or an edge case is not just suboptimal or hard-to-read code; it is a potential weapon waiting to be deployed at scale. The PM’s prioritization now has to account for the build-up of “security debt.” Choosing to defer a robustness test or a privacy-preserving architectural decision now can compound risk that only surfaces as a large, headline-grabbing failure (Tallam, 2025).

Table 3: *AI-Specific Security Metrics*

Metric Category	Example Metrics	PM Usage
Data Leak Prevention	Blocked PII per 1K queries	Measure privacy guard effectiveness
Model Robustness	% successful adversarial attacks	Gate model releases
Anomaly Detection	MTTD abnormal patterns	Justify automation investment
Governance	Audit logging coverage	Compliance reporting

Building a Shared Security Mindset

The approach outlined necessitates a close partnership between product, engineering, security, and legal teams that have not historically always seen eye-to-eye. For the PM, this means the soft skills of making security a non-technical, relevant risk to business stakeholders, as well as explaining the commercial and reputational stakes of security to engineers. Success would require cultural shifts away from “security throws things over the wall” to “security is part of the wall, not added on” (Sidhpurwala et al., 2025).

In sum, the security-first, proactive methodology this article advocates for is not just a defensive imperative but an essential ingredient in the larger paradigm of responsible AI. It acknowledges that safety, fairness, transparency, and accountability are functions of robust security; they cannot be retrofitted or grafted on (Tallam, 2025). The strategic and tactical decisions the PM makes in how to resource threat modeling, push back on risky architectures, and what to measure for in terms of security are now the direct arbiters of whether these higher principles are operationalized or remain wishful.

Conclusion

The PM as the Architect of Trustworthy AI

Artificial intelligence in consumer products is one of the most impactful shifts in product development, and with it comes an entirely new threat landscape. This article has tried to shed light on the specific risks around AI and why the solution to this is not the business-as-usual in the PM role.

AI's unique security challenges – from data leakage to poisoning and model manipulation, among others – require us to redefine how we approach our PM jobs. The PM role needs to strategically lead the implementation of a threat-focused, PM-driven approach. The approach we are advocating for is one that brings the PM directly into the process of threat modeling, security-first design, and adversarial validation in a prescriptive way.

PMs need to proactively de-risk AI initiatives by adopting a dual stance of technical-advocacy, translating known or likely technical threats into real, tangible, business cases for security and investing in AI security, while at the same time reigning in the wild west of AI development, by giving developers and designers a clear, actionable blueprint for productizing the twin pillars of trustworthy AI – data integrity and model robustness.

Embedding these efforts into the development process is the most critical and important distinction between simply building cool and novel AI features and building consumer AI that is secure, defensible, and ultimately successful. Product Managers who embed AI security into the heart of their products cease to be the bottleneck for feature releases and become the champions for product security. By embracing and understanding security in the context of AI, PMs get to change the security discussion for AI from one that solely takes place between technical experts into one that becomes the cornerstone of the PM's product strategy for long-term, trustworthy AI product success.

References

1. Agunbiade, O. L. (2025). Strategic Investment Analysis in Emerging Markets: A Framework for Value Creation, Financial Resilience, and Sustainable Private Equity Performance in Sub-Saharan Africa.
2. Isaac, E. R. H. P., & Reno, J. (2023). AI product security: A primer for developers. *arXiv*. https://doi.org/10.48550/arxiv.2304.11087
3. Rangaraju, S. (2023). Secure by intelligence: Enhancing products with AI-driven security measures. *EPH - International Journal of Science and Engineering*, *9*(3), 36–41. https://doi.org/10.53555/ephi.v9i3.212
4. Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). Navigating AI cybersecurity: Evolving landscape and challenges. *Journal of Intelligent Learning Systems and Applications*, *16*(3), 155–174. https://doi.org/10.4236/jilsa.2024.163010
5. Orugboh, O. G., Omabuwa, O. G., & Taiwo, O. S. (2024). Predicting Neighborhood Gentrification and Resident Displacement Using Machine Learning on Real Estate, Business, and Social Datasets. *Journal of Social Sciences and Community Support*, *1*(2), 53-70.
6. Orugboh, O. G., Omabuwa, O. G., & Taiwo, O. S. (2025). Predicting Intra-Urban Migration and Slum Formation in Developing Megacities Using Machine Learning and Satellite Imagery. *Journal of Social Sciences and Community Support*, *2*(1), 69-90.
7. Sidhpurwala, H., Mallett, G., Fox, E., Bestavros, M., & Chen, H. (2025). Building trust: Foundations of security, safety, and transparency in AI. *AI Magazine*, *46*(2). https://doi.org/10.1002/aaai.70005
8. Tallam, K. (2025a). Security-first AI: Foundations for robust and trustworthy systems. *arXiv*. https://doi.org/10.48550/arxiv.2504.16110